

Source-to-source optimising transformations of Prolog programs based on abstract interpretation

François Gobert and Baudouin Le Charlier

Université Catholique de Louvain, Belgium,
{francois.gobert,baudouin.lecharlier}@uclouvain.be

This work is supported by the Belgian FNRS fund.

Abstract. Making a Prolog program more efficient by transforming its source code, without changing its operational semantics, is not an obvious task. It requires the user to have a clear understanding of how the Prolog compiler works, and in particular, of the effects of ‘impure’ features like the cut. The way a Prolog code is written - e.g., the order of clauses, the order of literals in a clause, the use of cuts or negations - influences its efficiency. Furthermore, different optimisation techniques may be redundant or conflicting when they are applied together, depending on the way a procedure is called - e.g., inserting cuts and enabling indexing. We present an optimiser, based on abstract interpretation, that automatically performs safe code transformations of Prolog procedures in the context of some class of input calls. The method is more effective if procedures are annotated with additional information about modes, types, sharing, number of solutions and the like. Thus the approach is similar to Mercury. It applies to any Prolog program, however.

Keywords. Abstract interpretation, automated optimisation, cut insertion, logic programs, Prolog, source-to-source program transformation, static analysis.

1 Introduction

Programming in Prolog allows us to write so-called *multidirectional* procedures, in the sense that the same code of a procedure can be used in more than one way (the arguments being either input data or output results). The undesirable consequence for multidirectionality is inefficiency (in terms of space utilisation and of execution time). This efficiency issue comes from the general execution model of Prolog, which is generally based on the Warren’s Abstract Machine (WAM, for short) [1, 14]. Answer substitutions are computed according to a depth-first search strategy with backtracking, where clauses are executed from top-to-bottom, and literals are executed from left-to-right inside a clause. Given the incompleteness of Prolog, some input-output patterns can loop. Also, Prolog uses a general algorithm for unification, with no restriction on the terms being unified, but most compilers do not perform the occur-check test during the unification. For efficiency reasons, some built-in procedures are not multidirectional (for example, arithmetic and comparison predicates). Negation as failure is sound only if it applies to a ground literal. Due to the incompleteness and

unsoundness of Prolog, not every ordering of clauses and literals is operationally correct, and the way a procedure is written greatly influences the search of its solutions, and then, the efficiency.

A solution for optimising multidirectional procedures is to generate specialised code for each particular use of the procedure. In the context of a directionality, one can try to find a more efficient ordering of clauses and literals, such that the program still remains operationally correct for that directionality. In Prolog, we can also try to insert cuts to prune the search tree without removing solutions. This can greatly reduce the size of the search tree, and improves the efficiency. Applying correct code transformations is not obvious and is tricky to be done manually, because it is very error-prone. This paper describes an optimiser based on abstract interpretation, which realizes this task automatically.

To illustrate the interest of specialising code, consider the multidirectional procedure `efface(X,T,TEff)`, which is the running example of [6]: `X` is an element of list `T`, and `TEff` is the list `T` without the first occurrence of `X` in `T`.

```
efface(X,[H|T],[H|TEff]) :- efface(X,T,TEff), not(X=H).
efface(X,[X|T],T).
```

This code can be used in several ways: either when every input argument is a ground term; or when inputs `X` and `T` are ground and `TEff` is a variable; or when inputs `X` and `TEff` are ground and `X` is a variable; or when inputs `X` and `TEff` are variables and `T` is ground; etc. Now, if we consider only the first directionality, then our optimiser will be able to automatically generate the following specialised code (the optimiser has checked that the procedure is deterministic for this directionality):

```
efface(X,[X|T],T) :- !.
efface(X,[H|T],[H|TEff]) :- efface(X,T,TEff).
```

The clauses are reordered, a cut has been inserted in the first clause, and the negation is removed. For all inputs satisfying the first directionality, the sequences of answer substitutions of the specialised and of the multidirectional codes are identical. Table 1 compares the execution between the multidirectional and the specialised codes. Several tests have been performed by varying the list-length of the input list `T`. The table shows that the multidirectional code is less efficient than the specialised one in terms of execution time and of used local stack. In particular, the specialised code uses a constant amount of local stack (independently of the size of the input), while we yield a local stack error if we try to execute the multidirectional code with an input list of size 25000. The speedup increases according to the size of the input: for instance, the optimised code spent 3.61 times less execution time for an input list of size 10000.

Our approach is strongly inspired by [6], where a methodology to build correct programs is proposed: starting from a specification and a so-called *logic description* of the problem, the methodology constructs operationally correct programs which are not written in the usual style of experienced Prolog programmers: procedures are normalised, with explicit unifications, and are thus

Input list size	Execution Time (ms)			Used Local Stack (By)	
	Multidirectional (ms)	Optimized (ms)	(sdup)	Multidirectional (By)	Optimized (By)
100	71	25	2.84	8508	108
1000	693	225	3.09	84108	108
10000	8176	2264	3.61	840108	108
25000	ERROR	5633	-	ERROR	108

Table 1. Program `efface(X,T,TEff)` executed 1000 times, when `X` is a ground term, `T` is a ground list, and `TEff` is a variable. The program has been tested on a 1.5 Ghz Pentium; 1Gb RAM; Linux Suse; SWI-Prolog v 5.4.6 [15]. The size to which the local stack is allowed to grow is 2048000 By.

inefficient. The author of [6] then proposes to apply some code transformations, in order to produce more efficient programs (written in the usual style and where cuts are introduced). It is not obvious to ensure the correct application of transformations, nor to choose the best order to apply them. Our optimiser does not require or assume that Prolog programs are written in a specialised syntax. It accepts any kind of Prolog programs. So, the programmer has the liberty to write its program in the style he wants: normalised or not. Our optimiser then automatically specialises the program for some directionality, by choosing a suitable order for applying the transformations, and by ensuring that the transformations are correctly applicable.

The transformations performed by the system are related to partial evaluation: literals are evaluated at compile-time, such that some of them can be removed, and some unifications can be simplified. However, we do not unfold procedure calls as it is normally done in partial deduction.

The optimiser is based on an abstract interpretation framework [3, 10, 11], that collects and verifies the semantic information needed for correct application of source-to-source transformations. The operational properties caught and verified by the framework that are useful for the purpose of the optimiser are: cardinality information (including determinacy and conditions for sure success and failure), detection of the exclusivity between clauses, information about the mode, type, sharing, linearity, and size of input/output terms, occur-check freeness, and induction parameters for proving termination.

The rest of the paper is organised as follows. Section 2 presents the source-to-source transformations we apply on Prolog programs, and specifies the conditions for applying correctly such transformations. Section 3 describes briefly the abstract interpretation framework. Formal specifications are introduced, to allow the user to express the directionality for which the code must be optimised. Section 4 illustrates in which order the transformations are applied. Section 5 reports on experimental results, and Section 6 presents the related work.

2 Source-to-source Transformations

Optimisation is carried out by means of transformations based on the operational semantics of Prolog. Section 2.1 describes the Prolog programs accepted by the analyser, and the syntax of *normalised* programs, on which most code

transformations are applied. Section 2.2 introduces the concept of *sequence of answer substitutions*, that allows us to describe the sufficient conditions for applying correct source-to-source transformations. Section 2.3 defines some semantic properties that characterise sequences of answer substitutions. Such properties are undecidable, but can be safely approximated by our framework. Finally, Section 2.4 presents the transformation rules.

2.1 Prolog Syntax

The optimiser accepts any Prolog program that is ISO conformant, with special features like the cut and the negation, as well as constructs of the form $;$ (disjunctions) and $\rightarrow;$ (if then else). The abstract interpretation framework of the optimiser is designed on *normalised* programs, and most code transformations apply on such programs. A normalised procedure pr is a nonempty sequence of clauses c . Each normalised clause c has the form $h:-g$ where the head is of the form $p(X_1, \dots, X_n)$, whereas the body g is a possibly empty sequence of normalised literals. A normalised literal is either a built-in of the form $X_{i_1} = X_{i_2}$, a built-in of the form $X_{i_1} = f(X_{i_2}, \dots, X_{i_n})$, a procedure call $p(X_{i_1}, \dots, X_{i_n})$, a cut $!$, or a negation $not(l)$, where l is a normalised literal. The variables occurring in a literal are all distinct; all clauses of a procedure have exactly the same head; if a clause uses m different variables, these variables are X_1, \dots, X_m . Observe that all Prolog program can be rewritten into equivalent normalised programs. In the rest of this paper, P denotes the given normalised program.

2.2 Sequence of Answer Substitutions

Semantically, a normalised procedure pr can be viewed as a function mapping every input substitution θ to a sequence of answer substitutions S . A substitution θ is a finite set of the form $\{X_1/t_1, \dots, X_n/t_n\}$ where X_1, \dots, X_n are distinct program variables, and where t_1, \dots, t_n are terms (variables occurring in terms are *standard* variables; the sets of standard and program variables are disjoint). A sequence of answer substitutions S can be either finite $\langle \theta_1, \dots, \theta_k \rangle$ ($k \geq 0$), or infinite $\langle \theta_1, \dots, \theta_k, \dots \rangle$ ($k \in \mathbb{N}$), or incomplete $\langle \theta_1, \dots, \theta_k, \perp \rangle$ ($k \geq 0$), where the symbol \perp denotes that the procedure loops. To express this behaviour, we use the notation of [11]: $\langle \theta, pr \rangle \mapsto_P S$ for a procedure, $\langle \theta, c \rangle \mapsto_P S$ for a clause, and $\langle \theta, g, c \rangle \mapsto_P S$ for a prefix of a clause. In the rest of this paper, we assume that each procedure terminates. Thus, we only consider *finite* sequences of answer substitutions. Our optimiser is based on the abstract interpretation framework defined in [10], that uses induction parameters to verify that a procedure terminates. Notice that every procedure of our benchmarks terminates. The substitutions of S are denoted by $Subst(S)$, and the length of S is denoted by $|S|$.

2.3 Semantic Properties

This section defines some semantic properties that characterise the execution of literals, prefixes of the body of a clause, clauses, procedures, in terms of the length of their sequence of answer substitutions. Such properties are useful to express the sufficient conditions to apply safely the code transformations.

Let pr be a procedure of arity n , let c be a clause of the procedure pr , let (g, l) be a prefix of the clause c , where g is a goal and l is a literal. Let θ be an input program substitution with domain $\{X_1, \dots, X_n\}$. Consider the execution of the procedure $\langle \theta, pr \rangle \mapsto_P S_{pr}$, the execution of the clause $\langle \theta, c \rangle \mapsto_P S_c$, the execution of the prefix of the clause $\langle \theta, g, c \rangle \mapsto_P S_g$. Assume that the literal l is of the form $q(X_{i_1}, \dots, X_{i_r})$. The execution of the literal l after the execution of the goal g can be described for all $\theta' \in Subst(S_g)$ by $\langle \vartheta_{\theta'}, l \rangle \mapsto_P S_{\theta'}$, where $X_k \vartheta_{\theta'} = X_{i_k} \theta'$ ($1 \leq k \leq r$). We can now define the following properties (the terminology of [6] is used):

- The procedure pr , or the clause c , or the prefix of a clause g , or the literal after the execution of the goal g is **deterministic** w.r.t. θ iff their sequence of answer substitutions has *at most one* computed answer substitution: $|S_{pr}|, |S_c|, |S_g|, |S_{\theta'}| \in \{0, 1\}$, for all $\theta' \in Subst(S_g)$.
- The procedure pr , or the clause c , or the prefix of a clause g , or the literal after the execution of the goal g is **fully deterministic** w.r.t. θ iff their sequence of answer substitutions has *one and only one* answer substitution: $|S_{pr}| = |S_c| = |S_g| = |S_{\theta'}| = 1$, for all $\theta' \in Subst(S_g)$.
- The procedure pr , or the clause c , or the prefix of a clause g , or the literal after the execution of the goal g **surely succeeds** w.r.t. θ iff their sequence of answer substitutions has *at least one* answer substitution: $|S_{pr}|, |S_c|, |S_g|, |S_{\theta'}| \geq 1$, for all $\theta' \in Subst(S_g)$.
- The literal l is a **test literal** after the execution of the goal g w.r.t. θ if it is not a cut, and it is deterministic w.r.t. θ , and it does not instantiate any variable. For all $\theta' \in Subst(S_g)$, $S_{\theta'}$ is either the empty sequence $\langle \rangle$ or the sequence $\langle \vartheta_{\theta'} \rangle$.
- The two procedures pr_1 and pr_2 (with the same arity n) are **exclusive** w.r.t. θ iff either the execution of pr_1 fails, or the execution of pr_2 fails, or both executions of pr_1 and pr_2 fail:

$$\left. \begin{array}{l} \langle \theta, pr_1 \rangle \mapsto_P S_1 \\ \langle \theta, pr_2 \rangle \mapsto_P S_2 \end{array} \right\} \Rightarrow S_1 = \langle \rangle \text{ or } S_2 = \langle \rangle$$

The same definition applies for exclusivity between two clauses of same arity, or between two prefixes of clauses of same arity.

The above properties are undecidable but can be safely approximated by our abstract interpretation framework presented in Section 3.

2.4 Transformations Rules

The subsequent transformation rules are adapted from [6]. Let pr be a normalised procedures whose arity is n , and let θ be a program substitution whose domain is $\{X_1, \dots, X_n\}$. A rule transforming the procedure pr of the program P into the procedure pr' is *correct* w.r.t. θ if the executions of pr (in the context of the program P) and of pr' (in the context of the program P' , which is the program P where pr has been replaced by pr') produce the same sequence of answer substitutions. In other words, if $\langle \theta, pr \rangle \mapsto_P S$ and $\langle \theta, pr' \rangle \mapsto_{P'} S'$ then

$S = S'$. The conditions given for applying the transformations are expressed using the semantic properties defined in Section 2.3. Such conditions are sufficient but not always necessary. It is assumed that the procedure pr has no side-effects when it is executed with input θ .

Rule 1: Reorder clauses

$$\begin{array}{c}
\dots \\
c_i : p(X_1, \dots, X_n) :- g_i. \\
\dots \\
c_j : p(X_1, \dots, X_n) :- g_j. \\
\dots \\
\hline
\dots \\
\dots \\
c_j : p(X_1, \dots, X_n) :- g_j. \\
\dots \\
\dots \\
c_i : p(X_1, \dots, X_n) :- g_i. \\
\dots
\end{array}$$

where:

- for all $k \in \{i, \dots, j\}$, the clause c_k is deterministic w.r.t. θ ;
- for all $k, l \in \{i, \dots, j\} : k \neq l$, we have that c_k and c_l are exclusive w.r.t. θ ;
- for all $k \in \{i, \dots, j\}$, there is no cut in g_k .

Rule 2: Insert green cuts

$$\begin{array}{c}
c_1 : p(X_1, \dots, X_n) :- g_1. \\
\dots \\
c_k : p(X_1, \dots, X_n) :- l_1, \dots, l_i, l_{i+1}, \dots, l_s. \\
\dots \\
c_r : p(X_1, \dots, X_n) :- g_r. \\
\hline
c_1 : p(X_1, \dots, X_n) :- g_1. \\
\dots \\
c_k : p(X_1, \dots, X_n) :- l_1, \dots, l_i, !, l_{i+1}, \dots, l_s. \\
\dots \\
c_r : p(X_1, \dots, X_n) :- g_r.
\end{array}$$

where:

- the goal (l_1, \dots, l_i) is deterministic w.r.t. θ ;
- for all $z \in \{k+1, \dots, r\}$, the goal (l_1, \dots, l_i) and clause c_z are exclusive w.r.t. θ .

Rule 3: Eliminate dead code

$$\begin{array}{c}
c_1 : p(X_1, \dots, X_n) :- g_1. \\
\dots \\
c_k : p(X_1, \dots, X_n) :- l_1, \dots, l_i, !, l_{i+1}, \dots, l_s. \\
\dots \\
c_r : p(X_1, \dots, X_n) :- g_r. \\
\hline
c_1 : p(X_1, \dots, X_n) :- g_1. \\
\dots \\
c_k : p(X_1, \dots, X_n) :- l_1, \dots, l_i, !, l_{i+1}, \dots, l_s.
\end{array}$$

where:

- the goal (l_1, \dots, l_i) surely succeeds w.r.t. θ .

Rule 4: Move backwards cut

$$\begin{array}{c}
\dots \\
c_k : p(X_1, \dots, X_n) :- l_1, \dots, l_{i-1}, !, l_i, \dots, l_s. \\
\dots \\
\hline
\dots \\
c_k : p(X_1, \dots, X_n) :- l_1, \dots, l_{i-1}, l_i, !, \dots, l_s. \\
\dots
\end{array}$$

where:

- l_i is fully deterministic after the execution of goal (l_1, \dots, l_{i-1}) w.r.t. θ .

Rule 5: Remove useless test literals

$$\begin{array}{c}
c_1 : p(X_1, \dots, X_n) :- l_1, \dots, l_{i-1}, l_i, l_{i+1}, \dots, l_s. \\
\dots \\
c_k : p(X_1, \dots, X_n) :- l_1, \dots, l_{i-1}, l_i, l_{i+1}, \dots, l_s. \\
\dots \\
\dots \\
\hline
c_r : p(X_1, \dots, X_n) :- g_r. \\
c_1 : p(X_1, \dots, X_n) :- l_1, \dots, l_{i-1}, l_i, l_{i+1}, \dots, l_s. \\
\dots \\
c_k : p(X_1, \dots, X_n) :- l_1, \dots, l_{i-1}, l_{i+1}, \dots, l_s. \\
\dots \\
\dots \\
c_r : p(X_1, \dots, X_n) :- g_r.
\end{array}$$

where:

- l_i is a test literal after the execution of goal (l_1, \dots, l_{i-1}) w.r.t. θ ;
 - l_i is deterministic after the execution of goal (l_1, \dots, l_{i-1}) w.r.t. θ ;
 - l_i is fully deterministic after the execution of goal (l_1, \dots, l_{i-1}) w.r.t. θ
- or**
- $\exists z \in \{1, \dots, k-1\}$ such that a cut is surely executed in clause c_z w.r.t. θ .

The abstract interpretation framework that safely approximates the conditions for applying the above source-to-source transformations is discussed in the next section.

3 Abstract Interpretation Framework

This section presents the abstract interpretation framework [7, 10] that captures and checks the semantic properties (and other ones) defined in Section 2.3. This semantic information (e.g., about the determinacy, the exclusivity) is needed for ensuring correct application of the code transformations. Section 3.1 describes the two fundamental abstract domains. An *abstract substitution* represents a set of program substitutions, and an *abstract sequence* represents a set of sequences of answer substitutions. Section 3.2 illustrates the syntax of *formal specifications*, that allows the user to write abstract sequences into a convenient syntax. The abstract execution is briefly presented in Section 3.3. Finally, Section 3.4 shows how abstract domains are used to check the semantic properties.

3.1 Abstract Substitutions and Abstract Sequences

The domain of *abstract substitutions* is an instantiation to modes, types, linearity and possible sharing of the generic abstract domain $\mathbf{Pat}(\mathfrak{K})$ described in [4, 9]. An abstract substitution represents a set of program substitutions of the form $\{X_1/t_1, \dots, X_n/t_n\}$, where the X_i 's are program variables, and the t_i 's are terms (variables occurring in terms are standard variables; the set of standard and program variables are disjoint). The set of substitutions represented by β is denoted by $Cc(\beta)$. Formally, an abstract substitution β is a triple of the form $\langle sv, frm, \alpha \rangle$. The *same-value* (sv) and *frame* (frm) components provide information about the structure of terms. Each term described in β is represented by an index. The sv component maps each program variable X to its corresponding index. Hence, the equality $sv(X) = sv(Y)$ means that variables X and Y are bound to the same term. The frm component describes the pattern of some indices, by giving their functor name and the indices of their composing subterms. The *alpha tuple* α is the generic part of the domain. It provides extra information about all terms and subterms of interest (represented by the indices). In the current analyser, α is of the form $\langle mo, ty, ps, lin, E \rangle$. The mo component [9] maps each index to its mode (e.g., **gr** (ground), **var** (variable)). The ty component maps each index to its so-called *type expression* [7] (e.g., **list(int)** denotes the set of all lists of integers, **list(any)** denotes the set of all possibly non-instantiated lists). The ps component [9] is a binary relation over indices, and expresses the possible sharing between two terms. Pairs of indices that do not belong to ps surely do not share a variable. The lin component contains all indices that are surely linear (i.e., they do not contain several occurrences of the same variable). The E component is a set of linear relations between the size of terms (several norms can be combined). The abstract substitution whose concretisation is the empty set is denoted by \perp . The greatest lower bound between two abstract substitutions β_1 and β_2 is denoted by $\beta_1 \sqcap \beta_2$ and is such that $Cc(\beta_1) \cap Cc(\beta_2) = Cc(\beta_1 \sqcap \beta_2)$.

The domain of *abstract sequences* models the operational behaviour of a Prolog procedure. An abstract sequence B describes a set of pairs $\langle \theta, S \rangle$ where θ is a program substitution and S is the sequence of answer substitutions resulting from executing a procedure (a clause, a goal, etc.) with input substitution θ . An abstract sequence B is a tuple of the form $\langle \beta_{in}, \beta_{ref}, \beta_{fails}, U, \beta_{out}, E_{ref_out}, E_{sol} \rangle$ that imposes conditions on the pairs $\langle \theta, S \rangle$. The set of pairs $\langle \theta, S \rangle$ satisfying the conditions imposed by B are denoted by $Cc(B)$. The *input* abstract substitution β_{in} describes the class of accepted input calls: $\theta \in Cc(\beta_{in})$. The *refined* abstract substitution β_{ref} describes the successful input calls, i.e., those that produce at least one solution: $S \neq \langle \rangle$ implies $\theta \in Cc(\beta_{ref})$. The set of abstract substitutions β_{fails} describes conditions of sure failure: the sequence S is empty if the input substitution satisfies one of the abstract substitution of β_{fails} , i.e., if there exists $\beta_f \in \beta_{fails}$ such that $\theta \in Cc(\beta_f)$ then $S = \langle \rangle$. The *untouched* component U describes the set of input terms that are untouched (non-instantiated) during the execution. The *output* abstract substitution β_{out} describes the substitutions belonging to S : for each θ' in $Subst(S)$, we have that $\theta' \in Cc(\beta_{out})$.

The *size relations* component E_{ref_out} is a set of linear relations (equations and inequations) between the size of the input/output terms. The *cardinality* component E_{sol} is a set of linear relations between the number of solutions and the size of the input terms, i.e., $sol=|S|$ is a solution of the system E_{sol} .

3.2 Formal Specifications

Formal specifications describe abstract sequences using a concrete syntax more convenient for a programmer than the mathematical formalism of abstract sequences. A formal specification may contain other information needed for the analyser, like the induction parameter for proving termination. For instance, the following two specifications for the **efface** procedure can be written by the user (and can be checked by the analyser):

<pre> efface in(X:gr,T:list(gr),TEff:var) out(_,_,list(gr)) srel(TEff_out = T_in-1) sol(sol =< 1) sexpr(T) </pre>	<pre> efface in(X:gr,T:any,TEff:list(gr)) out(_,list(gr),_) srel(TEff_in = T_out-1) sol(sol =< TEff_in+1) sexpr(TEff) </pre>
---	--

We find the different parts of an abstract sequence. The first specification considers the situation where input **X** is a ground term, input **T** is a ground list, and **TEff** is a variable. After success of execution, **TEff** becomes a ground list, whose list-length is the list-length of input **T** minus one. The symbol ‘_’ is used when we do not provide refined information about an argument. The execution terminates (the size expression **T** decreases through recursive calls) and is deterministic ($sol \leq 1$). In the second specification, input **X** is a ground term, **T** is any term, and **TEff** is a ground list. This execution terminates and is non-deterministic (the number of solutions is between 0 and the list-length of input **TEff** plus one).

3.3 Abstract Execution and Annotated Procedures

Abstract execution is performed on normalised procedures (see Section 2.1), because it simplifies the design of abstract operations. The analyser then first translates a general Prolog procedure into an equivalent normalised procedure. The analysis is compositional. The system verifies a procedure against a specification, by assuming that the specifications hold for subproblems. For a given program, it analyses each procedure; for a given procedure, it analyses each clause; for a given clause, it analyses each atom. If an atom in the body of a clause is a procedure call, the analyser looks at the given specifications to infer information about its execution. The analyser succeeds if, for each procedure and each specification describing this procedure, the analysis of the procedure yields results that are covered by the considered specification.

As a result of the analysis, the procedure is annotated with abstract sequences at each program point. The annotation of the clause $c ::= p(X_1, \dots, X_n) :- l_1, \dots, l_s$. in the context of an abstract sequence $B_p = \langle \beta_{in}^p, \dots \rangle$ is of the form:

$$(\beta_{in}^p)p(X_1, \dots, X_n) :- (B_0)l_1, (B_1)\dots, l_s(B_s).(B_c)$$

The analyser certifies that every abstract sequence at a program point safely approximates the sequence of answer substitutions computed until that point. Let θ be an input program substitution in $Cc(\beta_{in}^p)$. For each program point i ($0 \leq i \leq s$), the concrete execution of l_1, \dots, l_i is approximated by B_i , i.e., $\langle \theta, (l_1, \dots, l_i), c \rangle \mapsto_P S_i$ implies $\langle \theta, S_i \rangle \in Cc(B_i)$. Similarly, the whole clause execution is approximated by B_c , i.e., $\langle \theta, c \rangle \mapsto_P S$ implies $\langle \theta, S \rangle \in Cc(B_c)$. The annotation of a procedure in the context of an abstract sequence B_p is the sequence of its annotated clauses.

3.4 Checking Semantic Properties

This section explains how the analyser can check the semantic properties of Section 2.3 that are useful for applying the code transformations. The components of the abstract sequences provide constraints about the length of the computed answer substitutions.

Let $B = \langle \beta_{in}, \beta_{ref}, \beta_{fails}, U, \beta_{out}, E_{ref_out}, E_{sol} \rangle$ be the abstract sequence approximating the execution of a procedure, or of a clause, or of a prefix in a clause, or a literal l after the execution of a goal. Let $B_i = \langle \beta_{in}, \beta_{ref}^i, \beta_{fails}^i, U^i, \beta_{out}^i, E_{ref_out}^i, E_{sol}^i \rangle$ be the abstract sequence modelling the execution of the clause c_i , or the execution of some prefix of the clause c_i ($1 \leq i \leq 2$), where c_1 and c_2 have the same name and arity.

- If **deterministic**(B) returns true then $\langle \theta, S \rangle \in Cc(B)$ implies $|S| \leq 1$. The value of **deterministic**(B) is set to true if **sol**=<1 is a solution of E_{sol} .
- If **fully_deterministic**(B) returns true then $\langle \theta, S \rangle \in Cc(B)$ implies $|S| = 1$. The value of **fully_deterministic**(B) is set to true if $\beta_{in} = \beta_{ref}$, and $\beta_{fails} = \emptyset$, and **sol**=1 is a solution of E_{sol} .
- If **test_literal**(B) returns true then $\langle \theta, S \rangle \in Cc(B)$ implies $S = \langle \rangle$ or $S = \langle \theta \rangle$. The value of **test_literal**(B) is set to true if **sol**=<1 is a solution of E_{sol} and if the *untouched* component U contains all the indices of β_{ref} .
- Let $\langle \theta, S_1 \rangle \in Cc(B_1)$ and $\langle \theta, S_2 \rangle \in Cc(B_2)$. If **exclusive**(B_1, B_2) returns true then $S_1 = \langle \rangle$ or $S_2 = \langle \rangle$. The value of **exclusive**(B_1, B_2) is set to true if one of the three following conditions holds:
 - the *refined* components β_{ref}^1 and β_{ref}^2 are incompatible: $\beta_{ref}^1 \sqcap \beta_{ref}^2 = \perp$
 - or $\exists \beta_f^1 \in \beta_{fails}^1 : (\beta_{ref}^1 \sqcap \beta_{ref}^2) \leq \beta_f^1$
 - or $\exists \beta_f^2 \in \beta_{fails}^2 : (\beta_{ref}^1 \sqcap \beta_{ref}^2) \leq \beta_f^2$

The accuracy and the cooperation between the abstract domains allow the analyser to detect automatically whether the conditions are satisfied for applying code transformations. The next section discusses the transformation strategy realized by the optimiser.

4 A Strategy to Generate Specialised Code

In the derivation of a procedure, there is often more than one possible sequence of transformations, resulting in different procedures. Some heuristics must then be included within the automation of the derivation of procedures to find permutations leading to efficient procedures for a given specification. For instance, the following heuristics are suggested in [6]:

- *Choose tail recursive permutations.* Last call optimisation is implemented in the WAM [14], such that the environment is deallocated *before* executing the last call. The efficiency gain is substantial if the last literal is a recursive call.
- *Choose permutations of the literals with the longest deterministic prefix.* This choice prevents useless computed answer substitutions by prefixes of the literals. Multiple answer substitutions are only generated by the suffixes.
- *Choose permutations of the literals that support the introduction of cuts and the removal of useless literals.*
- *Choose permutations of the literals such that the unifications are at the beginning.* This is useful to instantiate the clause heads and to suppress these equality literals.

The combination of transformation rules performed by the optimiser is guided by the above heuristics, and is illustrated on the example **efface**, whose initial multidirectional source code is:

```
efface(X,[H|T],[H|TEff]) :- efface(X,T,TEff), not(X=H).
efface(X,[X|T],T).
```

The following directionality is considered (expressed into a formal specification):

```
efface
in(X:gr, T:list(gr), TEff:any)
sol(sol =< 1)
```

STEP A: Syntactic normalisation. Most code transformations apply on normalised programs, and the abstract execution itself is defined on normalised programs, because it facilitates the analysis. Thus, the first step of the optimiser consists of translating the original procedure into an equivalent normalised code:

```
efface(X1,X2,X3) :- X2=[X4|X5], X3=[X4|X6], efface(X1,X5,X6), not(X1=X4).
efface(X1,X2,X3) :- X2=[X1|X3].
```

STEP B: Code annotation. In this step, the information needed to apply the code transformations is captured by the checker. Every clause of the normalised procedure is annotated with abstract sequences at each program point.

STEP C: Clause reordering. If the order of solutions is not modified, some clause reordering is achieved (**Rule 1**): a clause containing a literal that is a good candidate to be removed after the possible introduction of cuts, is placed at the bottom of the procedure. In the example **efface** and for the considered directionality, the following code is generated (clause reordering can be realized because the procedure is deterministic):

```
efface(X1,X2,X3) :- X2=[X1|X3].
efface(X1,X2,X3) :- X2=[X4|X5], X3=[X4|X6], efface(X1,X5,X6), not(X1=X4).
```

STEP D: Semantic normalisation. In order to insert a cut at the *leftmost* position in a clause, it may be useful to decompose a unification that may fail into equivalent but simpler unifications. It may then happen that a cut will be placed between such unifications, instead of after the (unique) global unification. This step is called *semantic normalisation* to distinguish it with the *syntactic normalisation* performed at the step A: the optimiser uses the semantic information available at each program point about the structure, the mode, the type,

the sharing and the linearity of terms, as well as the sure success of the execution of a unification. The following code is generated for **efface** (the last clause is not normalised semantically because no cut will be inserted there):

```
efface(X1,X2,X3) :- X2=[X4|X5], X4=X1, X5=X3.
efface(X1,X2,X3) :- X2=[X4|X5], X3=[X4|X6], efface(X1,X5,X6), not(X1=X4).
```

In the first clause, the unification $X2=[X1|X3]$ has been decomposed in three elementary unifications: the initial unification succeeds if $X2$ is a non-empty list (i.e., $X2=[X4|X5]$), and if the first element of $X2$ is $X1$ (i.e., $X4=X1$), and if the tail of $X2$ is $X3$ (i.e., $X5=X3$).

STEP E: Leftmost cut insertion with dead code elimination. The objective of this step is to insert *leftmost* green cuts in every clause, except in the last clause (**Rule 2**). In our example, a green cut is introduced in the first clause, at the program point where it is exclusive with the second clause:

```
efface(X1,X2,X3) :- X2=[X4|X5], X4=X1, !, X5=X3.
efface(X1,X2,X3) :- X2=[X4|X5], X3=[X4|X6], efface(X1,X5,X6), not(X1=X4).
```

The cut has been inserted before the last unification $X5=X3$. If we had not performed the previous step, then the cut would have been placed at the end of the first clause. When inserting a cut, some successive clauses may become useless, such that they can be removed (**Rule 3**).

STEP F: Move cuts backwards. The objective of this step is to obtain the longest deterministic prefix before executing a cut in a clause. While making sure that the procedure still terminates and that the subcalls are still correctly moded and typed, some literals are reordered inside the clauses. An inserted cut is moved backwards by passing literals that surely succeed before the cut (**Rule 4**). In our example, the cut cannot be moved backwards, because the unification $X5=X3$ does not surely succeed (at that point, $X5$ is a ground list and $X3$ is any term).

STEP G: Removing useless literals. The analyser is able to capture the input conditions for which a cut is *surely* executed. This information is used to refine the input conditions of the successive clauses. This allows the optimiser to remove some literals which become useless (**Rule 5**) (e.g., negation, test predicates, arithmetic built-ins). In our example, the cut is surely executed when the first element of input list $X2$ is the input $X1$. The second clause is thus surely not executed for that input. In particular, the negation surely succeeds and can therefore be suppressed safely:

```
efface(X1,X2,X3) :- X2=[X4|X5], X4=X1, !, X5=X3.
efface(X1,X2,X3) :- X2=[X4|X5], X3=[X4|X6], efface(X1,X5,X6).
```

STEP H: Semantic denormalisation. The code generated until this step is still normalised (with possibly added cuts). Thus, it remains inefficient. The last step consists of applying the reverse transformation. The *semantic denormalisation* uses the information captured by the analyser to suppress explicit unifications, or to replace them by simpler ones, and to place them implicitly in the head of clauses. The specialised code for **efface** is thus finally generated:

```
efface(X1,[X1|X2],X3)      :- !, X2=X3.
efface(X1,[X4|X2],[X4|X3]) :- efface(X1,X2,X3).
```

Remark. It may happen that the initial source code already contains some cuts. In such situation, the optimiser first removes every green cut, such that only the necessary *leftmost* cuts will be introduced during the step E.

5 Experimental Evaluation

Table 2 compares the execution between the original and the generated versions of `efface(X,T,TEff)` presented in the previous section. Several tests have been performed by considering executions that succeed and that fail, and by varying the list-length of the input list `T`. The table shows that the multidirectional code is less efficient than the specialised one in terms of execution time and of used local stack. In particular, the specialised code uses a constant amount of local stack (independently of the size of the input), while we yield a local stack error if we try to execute the multidirectional code with an input list of size 25000.

Success of execution	Input list size	Execution Time (ms)			Used Local Stack (By)	
		Multidirectional (ms)	Optimized (ms)	(sdup)	Multidirectional (By)	Optimized (By)
no	100	41	23	1.83	8508	156
no	1000	380	194	1.96	84108	156
no	10000	4268	1893	2.25	840108	156
no	25000	ERROR	4715	-	ERROR	156
yes	100	115	26	4.39	8508	156
yes	1000	689	221	3.11	84108	156
yes	10000	7334	2207	3.32	840108	156
yes	25000	ERROR	5527	-	ERROR	156

Table 2. Program `efface(X,T,TEff)` executed 1000 times, when `X` is a ground term, `T` is a ground list, and `TEff` is any term. Several tests are performed, depending on whether the execution fails or succeeds, and depending on the list-length of `T`. The program has been tested on a 1.5 Ghz Pentium; 1Gb RAM; Linux Suse; SWI-Prolog v 5.4.6 [15]. The size to which the local stack is allowed to grow is 2048000 By.

The optimiser has been tested on some classical programs, borrowed from [2,6,13] and from the Internet. The source programs, the formal specifications, the generated specialised codes, and the efficiency tests are available at <http://www.info.ucl.ac.be/~gobert>. The tests have been realized on a 1.5 Ghz Pentium, 1Gb RAM, Linux Suse, with SWI-Prolog v 5.4.6 [15].

Tests on Execution Time. Table 3 reports on execution time speedup, defined as the ratio between the execution time spent for the source program and for the specialised program. A speedup greater than (resp. less than) one means that the specialised code is more (resp. less) efficient than the source code. We consider 59 procedures and 112 specifications (some procedures have several specifications, because they are multidirectional). The benchmark *all* is composed of 173 efficiency tests (there is at least one efficiency test for each specification of a procedure). The benchmark *det* is a subset of the benchmark *all*. It contains only the efficiency tests for the deterministic procedures. The benchmark *ss* contains the efficiency tests for the procedures that surely succeed. The benchmark *det+ss* contains the efficiency tests for the deterministic procedures

that surely succeed. The mean speedup ranges from 1.42 to 1.68. The maximal speedup is 8.54 and the minimum speedup is 0.59.

Efficiency tests	Execution Time (Sdup)		
	Mean	Max	Min
173 <i>all</i>	1.42	8.54	0.59
125 <i>det</i>	1.45	8.54	0.59
45 <i>ss</i>	1.6	8.54	0.8
38 <i>det+ss</i>	1.68	8.54	0.8

Table 3. Execution time speedup between source codes and specialized codes generated by the optimiser: 173 efficiency tests distributed out of 59 predicates and 112 formal specifications. From the 173 tests, we obtain a speed up for 112 tests, and a speed down for 61 tests.

Tests on Space Utilisation. Table 4 reports on local stack utilisation. 57 generated procedures are considered. The maximal amount of local stack used during the execution of the generated code is either reduced (for 28 procedures), or identical (17 procedures), or increased (12 procedures) w.r.t. the maximal amount of local stack used during the execution of the source code.

Generated procedures	Used local stack is reduced wrt source code	Used local stack is identical wrt source code	Used local stack is increased wrt source code
57	28	17	12

Table 4. Comparisons between source & generated codes in terms of space utilization.

Accuracy of the Tests. The results reported on Table 3 and Table 4 depend on the choice of the efficiency tests. It is impossible to perform tests that include every situation in the context of a directionality. We have tried to make sufficiently general tests for each directionality. The efficiency results depend also on the way the source code is written. For all the benchmark reported in the tables, the original program was written in the usual Prolog style (i.e., not normalised). The optimiser can take as input programs that are normalised, like the ones that are derived in the methodology [6]. If we perform the same efficiency tests with normalised programs as initial source code, then we obtain a mean speedup of 3. This shows the utility of the optimiser.

Generated code may sometimes be less efficient. In general, the optimiser generates code that is more efficient than the source code, in terms of execution time and of local stack utilisation. But the tables report that a generated code can sometimes be less efficient than the original code. For instance, this occurs with the `append(L1,L2,L3)` procedure which concatenates two lists (we consider the usual directionality where inputs L1 and L2 are ground lists and L3 is a variable). The initial source code is:

```
append([],L,L).
append([H|L1],L2,[H|L3]) :- append(L1,L2,L3).
```

In SWI-Prolog, indexing is enabled by default on the first argument. Thus, in the context of the considered directionality, no choice point is created on the local stack. Indeed, the right clause is directly selected according to the principal

functor of the first input list, which is either an empty or a non-empty list. Furthermore, the second clause is a *chain rule* (i.e., there is only one atom in the body), such that no environment frame is allocated on the local stack. Therefore, this source code uses a constant amount of local stack, and executes very quickly.

The optimiser does not take into account the indexing technique in its strategy for applying correct code transformations, and the following code will be generated:

```
append([H|L1],L2,[H|L3]) :- append(L1,L2,L3), !.
append(_,L,L).
```

where the two clauses are reordered, a cut is inserted after the recursive call, and the constant [] in the second clause is removed. This code is less efficient than the initial source code. Indexing has no effect because the right clause cannot be selected according to the principal functor of the input list L1. Thus, Prolog creates a choice point on the local stack each time the first clause is called (and it is often executed because it is the recursive call). The cut is a *deep* cut (i.e., it is not located just after the symbol :-). It occurs after the recursive call, such that the first clause is no more a *chain rule*, and an environment frame must be allocated on the local stack each time we execute the first clause. The amount of local stack used during execution is increasing through recursive calls, to the contrary of the initial source code, which uses a constant amount of local stack.

6 Related Work

The Mercury programming language [12] is associated with a whole range of analysis tools for optimisation purposes. In Mercury also, the programmer has to annotate the program with information about modes, types, success and determinacy. A main difference is that not all logic programs are accepted by Mercury (only limited forms of unification are allowed). There are restrictions on the form of the programs and queries in order to generate more efficient code. Mercury is not based on the Warren's Abstract Machine, but has specialised - more efficient - algorithms, depending on the determinacy information. So our approach is more appealing to programmers who are willing to keep the full power of Prolog. Unlike Mercury, we do not change the usual execution model of Prolog based on the WAM [1, 14]: our optimiser performs some transformations at the Prolog code level. Actually, our optimiser should be able to generate low-level code. But performing source-to-source transformations is more portable (and easier to explain) than generating specialised WAM's instructions. Most Mercury annotations can be translated into our formal specification language. A current limitation to our language is that, for instance, we cannot express that an input list contains only free distinct variables (this can be expressed in Mercury). On the other hand, more general directionalities can be described in our language. For instance, we can express that some argument is a list possibly non-instantiated and possibly non-linear, and that two terms possibly share a variable (this cannot be expressed in Mercury), like in the following formal specification for `append`:

```
append
  in(L1:list(any), L2:list(any), L3:any)
  out(_, _, list(any))
  sol(sol =< 1)
```

The Ciao preprocessor CiaoPP [8] is a powerful static analyser based on abstract interpretation, which features many analyses similar to ours and other ones. The system can infer and/or check properties like regular types, modes, sharing, non-failure and determinacy, bounds on computational cost, bounds on sizes of terms in the program, and termination. In that system, procedures can be optionally annotated by assertions, which partially corresponds to specifications of our system. The system can perform automatic optimisations such as source-to-source transformation, specialisation, partial evaluation of programs, program parallelisation. Some transformations like cut insertion and semantic denormalisation are not performed in CiaoPP.

The authors of [5] consider how most of the common uses of cut can be eliminated from Prolog source programs, by relying on static analysis to generate them at compile time. Static analysis techniques are used to detect situations where to place cuts. In our approach, the insertion of cuts is only one part of the optimisation process: several source-to-source transformations are applied to find the best place where to place the cut (e.g., clause and literal reordering, semantic normalisation, etc.), to remove literals becoming useless due to the execution of cut in previous clauses, and to perform some partial evaluation.

7 Conclusion and Future Work

We have presented an optimiser based on abstract interpretation which attempts to make a Prolog program more efficient by transforming its source code without changing its operational semantics. The tool automatically performs safe code transformations of any Prolog program in the context of some class of input calls (described in formal specifications). Preliminary experimental tests of the optimiser show encouraging results, since the specialised codes are, at the average, 1.42 time more efficient in terms of execution time consumption and of local stack utilisation. The optimiser can be improved: we plan to find better heuristics for specialising the code, based on the knowledge of the WAM [1, 14], in order to take into account the indexing technique, and the influence on the efficiency of adding or not *deep* cuts.

References

1. H. Ait-Kaci. *Warren's Abstract Machine: A Tutorial Reconstruction*. 1991. The MIT Press. London.
2. K.R. Apt. *From Logic Programming to Prolog*. Prentice Hall, 1997.
3. Bruynooghe, M. 1991. *A practical framework for the abstract interpretation of logic programs*. J. Log. Program. 10, 2 (Jan. 1991), 91-124.
4. A. Cortesi, B. Le Charlier, P. Van Hentenryck. *Combination of abstract domains for logic programming: open product and generic pattern construction*. Science of Computer Programming 38:27-71, Elsevier Science 2000.
5. S. K. Debray and D. S. Warren. *Towards Banishing the Cut from Prolog*. IEEE Trans. Softw. Eng. 16, 3 (Mar. 1990), 335-349. 1990.
6. Y. Deville. *Logic Programming: Systematic Program Development*. A. Wesley, 1990.
7. F. Gobert and B. Le Charlier. *A System to check Operational Properties of Logic Programs*. AFADL'07 conference. Namur. 2007.

8. M. Hermenegildo, G. Puebla, F. Bueno, P. López-García. *Integrated Program Debugging, Verification, and Optimization Using Abstract Interpretation (and The Ciao System Preprocessor)*. Science of Computer Programming, Vol. 58, Num. 1-2, pages 115-140, Elsevier Science, October 2005.
9. B. Le Charlier and P. Van Hentenryck. *Experimental Evaluation of a Generic Abstract Interpretation Algorithm for Prolog*. ACM Transactions on Programming Languages and Systems (TOPLAS), 16(1):35-101, January 1994.
10. B. Le Charlier, C. Leclère, S. Rossi, and A. Cortesi. *Automated Verification of Prolog Programs*. JLP, 39:3-42, Elsevier Science 1999.
11. B. Le Charlier, S. Rossi, and P. Van Hentenryck. *Sequence-Based Abstract Interpretation of Prolog*. TPLP 2(1): 25-84 (2002)
12. Z.Somogyi, F.Henderson, T.Conway. *The Execution Algorithm of Mercury, an Efficient Purely Declarative Logic Programming Language* JLP, 29(1-3):17-64, 1996.
13. L. Sterling and E. Shapiro. *The Art of Prolog, Advanced Programming Techniques*. Second Edition. The MIT Press. 1994. Cambridge, Massachusetts. London.
14. D. H. D. Warren. *An abstract Prolog instruction set*. Technical Note 309, SRI International, Menlo Park, CA, October 1983.
15. J. Wielemaker. *SWI-Prolog 5.6 Reference Manual*. University of Amsterdam. April 2006. <http://www.swi-prolog.org>